AllLife Bank

Project 2 – AIML for Business

Matthew Clark - April 20, 2024

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Executive Summary

- AllLife Banks would like to quickly grow the number of loan customers in order to increase bank revenue. One way to achieve this, based on past marketing, is to convert current customers. The Data Science team was able to analyze and model the database of current customes to identify the customers that have the highest probability of purchasing loans. With that information we can make several actionable recommendations.
- Our modeling indicates that Income is the most important factor when identifying the current customers that are most likely to purchase a personal loan.
- Marketing should be primarily focused on current customers that have more than 98.5 in the Income category to achieve their goal of quickly increasing loan customers.
- The data also shows that customers with CD accounts are much more likely to get a personal loan so marketing to CD account holders is advised.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Executive Summary (continued)

- There is a strong correlation between Mortgages and Income. Mortgage holders should also be marketed too as they may have need for loan services to make improvements to their homes.
- In addition, 60% of customers use the online services of the bank so we recommend heavily marketing loan products on the website.
- Credit Card use also correlates to personal loans. Loan products should be heavily marketed to customers that have bank credit cards.
- More educated customers have a higher income and should also be a primary target for marketing loan products.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Business Problem Overview and Solution Approach

- AllLife Banks currently has a growing base of liability customers (deposits) with a small number of asset customers (loans). The bank would like to rapidly increase their loan business. One way to achieve this goal is targeting existing liability customers and adding them as assets customers also. Past marketing has shown a strong conversion rate is possible.

- The Data Science team has been tasked with creating a model , using the database of current liability customers, to assist the retail marketing team in identifying the potential asset customers with the highest probabilty of purchasing a loan.

- Our approach to this task will start with processing and analyzing the dataset to ensure accuracy and best modeling. Next a Decision Tree model will be constructed and deployed using baseline, pre-pruning, post-pruning, and cost-complexity pruning. The output of the different models will be assessed and compared using recall as the primary metric to select the best model for this task.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results

- Age and Experience have even distributions and are overall very similar.
- Income overall is right skewed with some outliers. While the average income is 73,000, roughly half of customers make less than the average.
- When looking at customers that have personal loans with the bank the income data is left skewed. There is a strong correlation between income and customers that have personal loans.
- Customers that do not have personal loans have right skewed data and account for all of the outliers within income.
- Overall CCAvg is heavily right skewed and contains the largest number of outliers.
- There is a strong correlations between Income and CCAvg indicating customers with more income are more likely to have and use a bank credit card.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results

- 30% of bank customers have a credit card issued by the bank.
- Customers with a personal loan spend more on credit cards each month.
- Mortgage is highly right skewed and contains the second most outliers. Mortgage has its stongest correlation with income.
- Roughly 65% of bank customers do no currently have a mortgage.
- The number of family members is pretty evenly spread from 1 to 4 among customers.
- More than 60% of bank customers have graduate or advanced degree. Customers with graduate and advanced degrees are more likely to have a personal loan.
- Only 10% of customers have a Securities Account. Only 6% of customers have a CD Account. But 50% of personal loan customers also have a CD Account.
- Nearly 60% of customers used the banks online services.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Preprocessing

- No duplicate values were identified.
- No missing values were detected.
- Anomalous values were found in Experience and corrected.
- The ID Column was removed because it provided no unique information.
- The Experience column was removed because it perfectly correlates with Age.
- Outliers were found in Income (1.92%), Mortgage (5.82%), and CCAvg (6.48%)
- Because the percentages of outliers were relatively low, and the categories could all reasonably be found to contain explainable outliers they were not treated.
- The ZIPcode column was treated to reduce the number of unique entries from 467 to 7.
- Categorical data columns were converted to ‘category’.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Building

- Data is next prepared for modeling by dropping the Personal Loan and Experience columns and splitting the data 70/30 between train and test with random_state1.
- To begin preparing the model we create sk_learn functions to plot confusion matrix and evaluate the performance of each model. This will allow easy evaluation for different model versions.
- Now we build the original Decision Tree Model.
- The original model paramaters were criterion='gini', random_state=1

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Improvement

- The scoring for the original model indicates overfitting so we will next apply pruning techniques to improve the model.
- For pre-pruning we used Grid Search CV to identify the parameters for best fit. Grid Search CV identfied the parameters as max-depth = 6, max-leaf_nodes=10, random_state=1
- The pre-pruned model did not score well so we move on to post-pruning.
- We used cost-complexity pruning to identify the ccp_alpha value to be used in post pruning.
- Post-pruning parameters were ccp_alpha=0.4708834100596766, class_weight=[0; 0.15, 1: 0.85], random_state = 1
- The post-pruned model showed an improvement over the pre-pruned model.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- To evaluate each model, we are focused on recall scoring.
- The post-pruned model was much less complex that the original or pre-pruned models.
- The rules for the post-pruned model:

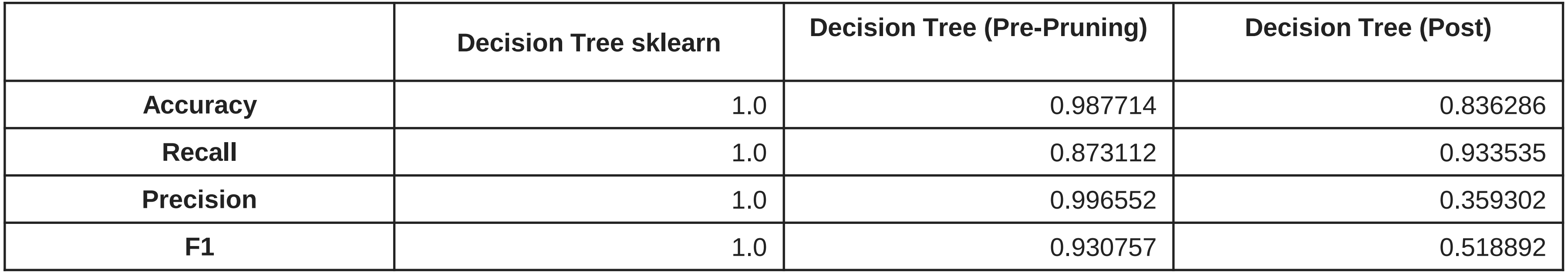
```
|--- Income <= 98.50
|   |--- weights: [392.70, 18.70] class: 0
|--- Income >  98.50
|   |--- weights: [82.65, 262.65] class: 1
```

- The post-pruned model gives all the feature importance to Income.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

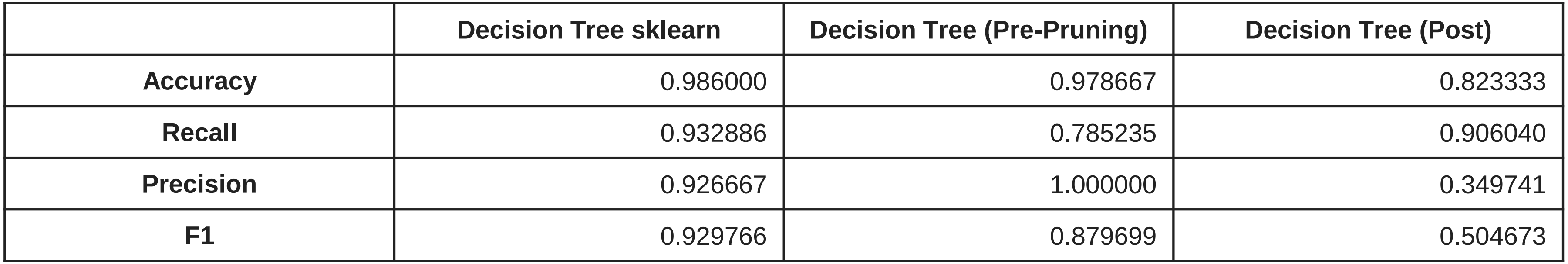
Model Performance Comparisons

- Training Performance Comparisons:



	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post)
Accuracy	1.0	0.987714	0.836286
Recall	1.0	0.873112	0.933535
Precision	1.0	0.996552	0.359302
F1	1.0	0.930757	0.518892

- Testing Performance Comparisons:



	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post)
Accuracy	0.986000	0.978667	0.823333
Recall	0.932886	0.785235	0.906040
Precision	0.926667	1.000000	0.349741
F1	0.929766	0.879699	0.504673

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

APPENDIX

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Background and Contents

Data Dictionary

- ID: Customer ID
- Age: Customer’s age in completed years
- Experience: #years of professional experience
- Income: Annual income of the customer (in thousand dollars)
- ZIP Code: Home Address ZIP code.
- Family: the Family size of the customer
- CCAvg: Average spending on credit cards per month (in thousand dollars)
- Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

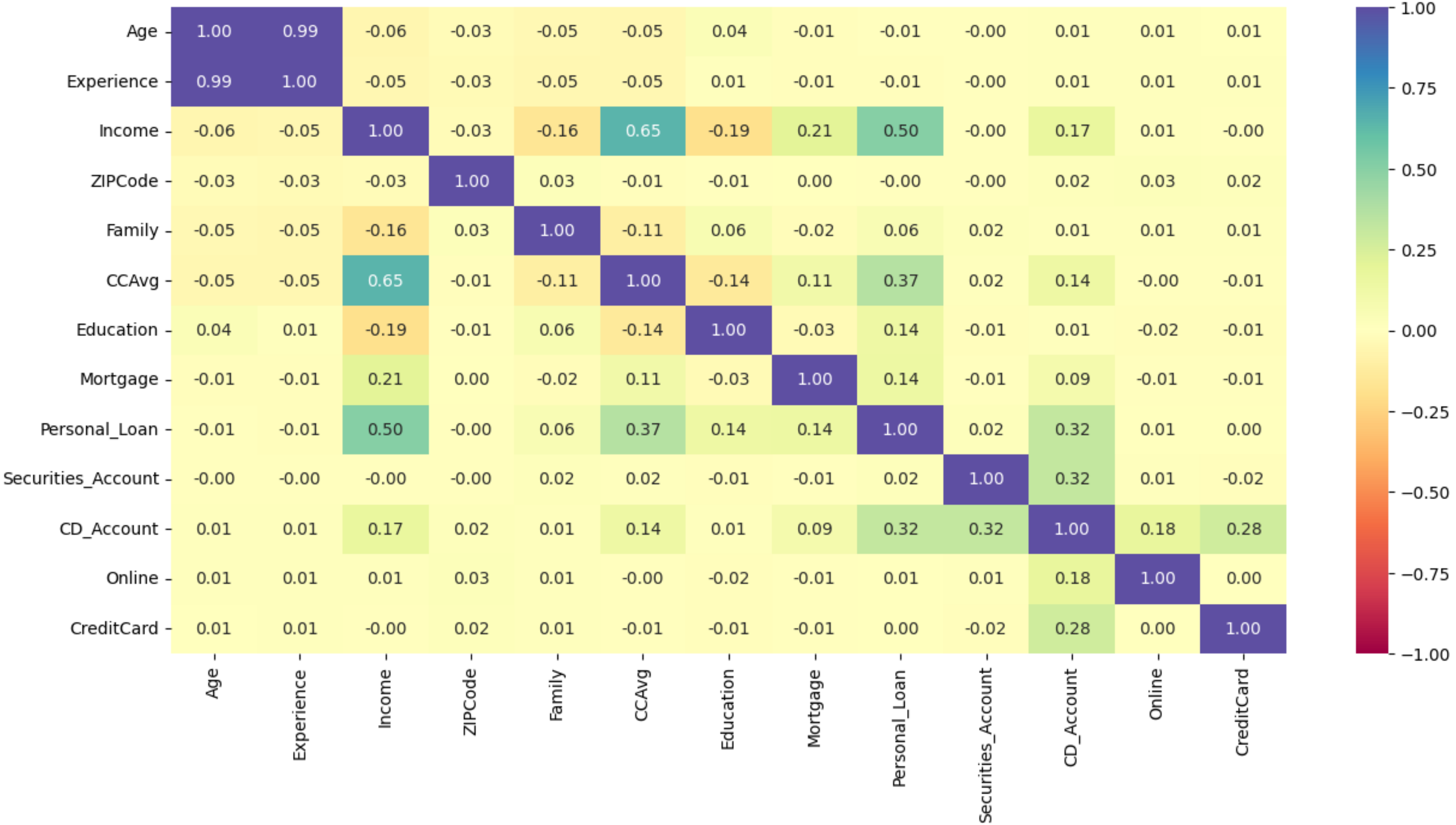
Data Background and Contents

Data Dictionary (cont.)

- Mortgage: Value of house mortgage if any. (in thousand dollars)
- Personal_Loan: Did this customer accept the personal loan offered in the last campaign? (0: No, 1: Yes)
- Securities_Account: Does the customer have securities account with the bank? (0: No, 1: Yes)
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank? (0: No, 1: Yes)
- Online: Do customers use internet banking facilities? (0: No, 1: Yes)
- CreditCard: Does the customer use a credit card issued by any other Bank (excluding All life Bank)? (0: No, 1: Yes)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

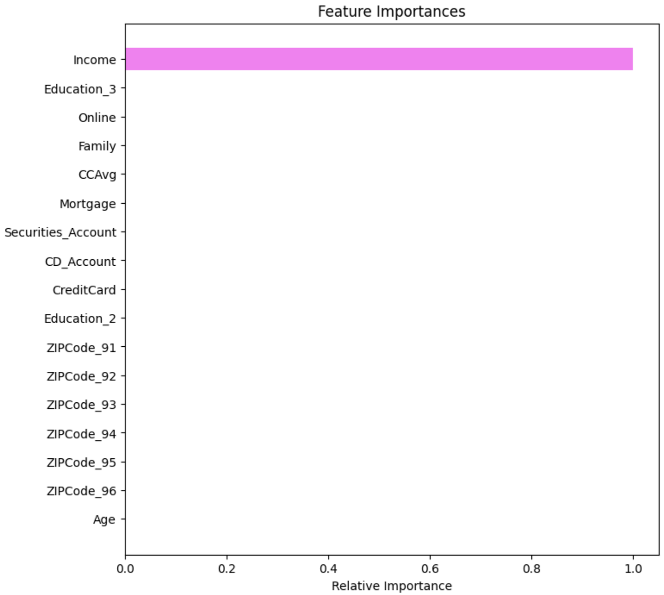
Data Correlations

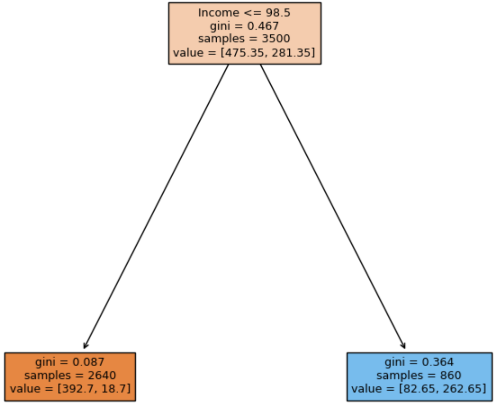


	Age	Experience	Income	ZIPCode	Family	CCAvg	Education	Mortgage	Personal_Loan	Securities_Account	CD_Account	Online	CreditCard
Age	1.00	0.99	-0.06	-0.03	-0.05	-0.05	0.04	-0.01	-0.01	-0.00	0.01	0.01	0.01
Experience	0.99	1.00	-0.05	-0.03	-0.05	-0.05	0.01	-0.01	-0.01	-0.00	0.01	0.01	0.01
Income	-0.06	-0.05	1.00	-0.03	-0.16	0.65	-0.19	0.21	0.50	-0.00	0.17	0.01	-0.00
ZIPCode	-0.03	-0.03	-0.03	1.00	0.03	-0.01	-0.01	0.00	-0.00	-0.00	0.02	0.03	0.02
Family	-0.05	-0.05	-0.16	0.03	1.00	-0.11	0.06	-0.02	0.06	0.02	0.01	0.01	0.01
CCAvg	-0.05	-0.05	0.65	-0.01	-0.11	1.00	-0.14	0.11	0.37	0.02	0.14	-0.00	-0.01
Education	0.04	0.01	-0.19	-0.01	0.06	-0.14	1.00	-0.03	0.14	-0.01	0.01	-0.02	-0.01
Mortgage	-0.01	-0.01	0.21	0.00	-0.02	0.11	-0.03	1.00	0.14	-0.01	0.09	-0.01	-0.01
Personal_Loan	-0.01	-0.01	0.50	-0.00	0.06	0.37	0.14	0.14	1.00	0.02	0.32	0.01	0.00
Securities_Account	-0.00	-0.00	-0.00	-0.00	0.02	0.02	-0.01	-0.01	0.02	1.00	0.32	0.01	-0.02
CD_Account	0.01	0.01	0.17	0.02	0.01	0.14	0.01	0.09	0.32	0.32	1.00	0.18	0.28
Online	0.01	0.01	0.01	0.03	0.01	-0.00	-0.02	-0.01	0.01	0.01	0.18	1.00	0.00
CreditCard	0.01	0.01	-0.00	0.02	0.01	-0.01	-0.01	-0.01	0.00	-0.02	0.28	0.00	1.00

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Post Pruned Model Visualizations





Income <= 98.5
gini = 0.467
samples = 3500
value = [475.35, 281.35]

gini = 0.087
samples = 2640
value = [392.7, 18.7]

gini = 0.364
samples = 860
value = [82.65, 262.65]

Happy Learning !